Computational information geometry in statistics: theory and practice

Entropy **2014**, *16*, 2454-2471; doi:10.3390/e16052454

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Computational Information Geometry in Statistics: Theory and Practice

Frank Critchley [1] and Paul Marriott [2,*]

[1] Department of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, UK; E-Mail: f.critchley@open.ac.uk

[2] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

* Author to whom correspondence should be addressed; E-Mail: pmarriot@uwaterloo.ca; Tel.: +1-519-888-4567.

Received: 27 March 2014; in revised form: 25 April 2014 / Accepted: 29 April 2014 / Published: 2 May 2014

Abstract: A broad view of the nature and potential of computational information geometry in statistics is offered. This new area suitably extends the manifold-based approach of classical information geometry to a simplicial setting, in order to obtain an operational universal model space. Additional underlying theory and illustrative real examples are presented. In the infinite-dimensional case, challenges inherent in this ambitious overall agenda are highlighted and promising new methodologies indicated.

Keywords: information geometry; computational geometry; statistical foundations

---

1. Introduction

The application of geometry to statistical theory and practice has seen a number of different approaches developed. One of the most important can be defined as starting with Efron's seminal paper [1] on statistical curvature and subsequent landmark references, including the book by Kass and Vos [2]. This approach, a major part of which has been called information geometry, continues today, a primary focus being invariant higher-order asymptotic expansions obtained through the use of differential geometry. A somewhat representative example of the type of result it generates is taken from [2], where the notation is defined:

Example 1 *The bias correction of a first-order efficient estimator, $\hat{\beta}$, is defined by:*

$$b^a(\beta) = -\frac{1}{2n} g^{aa'} \left\{ g^{bc} \Gamma^{(-1)}_{a'bc} + g^{\kappa\lambda} h^{(-1)}_{\kappa\lambda a'} \right\},$$

and has the property that if $\hat{\beta}^* := \hat{\beta} - b(\beta)$ then:

$$E_\beta(\hat{\beta}^* - \beta) = O(n^{-3/2}).$$

The strengths usually claimed of such a result are that, for a worker fluent in the language of information geometry, it is explicit, insightful as to the underlying structure and of clear utility in statistical practice. We agree entirely. However, the overwhelming evidence of the literature is that, while the benefits of such inferential improvements are widely acknowledged in principle, in practice, the overhead of first becoming fluent in information geometry prevents their routine use. As a result, a great number of powerful results of practical importance lay severely underused, locked away behind notational and conceptual bars.

This paper proposes that this problem can be addressed computationally by the development of what we call computational information geometry. This gives a mathematical and numerical computational framework in which the results of information geometry can be encoded as "black-box" numerical algorithms, allowing direct access to their power. Essentially, this works by exploiting the structural properties of information geometry, which are such that all formulae can be expressed in terms of four fundamental building blocks: defined and detailed in Amari [3], these are the $+1$ and -1 geometries, the way that these are connected via the Fisher information and the foundational duality theorem. Additionally, computational information geometry enables a range of methodologies and insights impossible without it; notably, those deriving from the operational, universal model space, which it affords; see, for example, [4–6].

The paper is structured as follows. Section 2 looks at the case of distributions on a finite number of categories where the extended multinomial family provides an exhaustive model underlying the corresponding information geometry. Since the aim is to produce a computational theory, a finite representation is the ultimate aim, making the results of this section of central importance. The paper also emphasises how the simplicial structures introduced here are foundational to a theory of computational information geometry. Being intrinsically constructive, a simplicial approach is useful both theoretically and computationally. Section 3 looks at how simplicial structures, defined for finite dimensions, can be extended to the infinite dimensional case.

2. Finite Discrete Case

2.1. Introduction

This section shows how the results of classical information geometry can be applied in a purely computational way. We emphasise that the framework developed here can be implemented in a purely algorithmic way, allowing direct access to a powerful information geometric theory of practical importance.

The key tool, as explained in [4], is the simplex:

$$\Delta^k := \left\{ \pi = (\pi_{\mathbf{0}}, \pi_{\mathbf{1}}, \ldots, \pi_{\mathbf{k}})^\top \,:\, \pi_{\mathbf{i}} \geq \mathbf{0}\,,\, \sum_{\mathbf{i=0}}^{\mathbf{k}} \pi_{\mathbf{i}} = \mathbf{1} \right\}, \tag{1}$$

with a label associated with each vertex. Here, k is chosen to be sufficiently large, so that any statistical model—by which we mean a sample space, a set of probability distributions and selected inference problem—can be embedded. The embedding is done in such a way that all the building blocks of information geometry (*i.e.*, manifold, affine connections and metric tensor) can be numerically computed explicitly. Within such a simplex, we can embed a large class of regular exponential families; see [6] for details. This class includes exponential family random graph models, logistic regression, log-linear and other models for categorical data analysis. Furthermore, the multinomial family on $k+1$ categories is naturally identified with the relative interior of this space, $int(\Delta^k)$, while the extended family, Equation (1), is a union of distributions with different support sets.

This paper builds on the theory of information geometry following that introduced by [3] via the affine space construction introduced by [7] and extended by [8]. Since this paper concentrates on categorical random variables, the following definitions are appropriate. Consider a finite set of disjoint categories or bins $\mathcal{B} = \{B_i\}_{i \in A}$. Any distribution over this finite set of categories is defined by a set, $\{\pi_i\}_{i \in A}$, which defines the corresponding probabilities. With "mix" connoting mixtures of distributions, we have:

Definition 1 *The -1-affine space structure over distributions on $\mathcal{B} := \{B_i\}_{i \in A}$ is $(X_{mix}, V_{mix}, +)$ where:*

$$X_{mix} = \left\{ \{x_i\}_{i \in A} | \sum_{i \in A} x_i = 1 \right\}, V_{mix} = \left\{ \{v_i\}_{i \in A} | \sum_{i \in A} v_i = 0 \right\}$$

and the addition operator, $+$, is the usual addition of sequences.

In Definition 1, the space of (discretised) distributions is a -1-convex subspace of the affine space, $(X_{mix}, V_{mix}, +)$. A similar affine structure for the $+1$-geometry, once the support has been fixed, can be derived from the definitions in [7].

2.2. Examples

Examples 2 and 3 are used for illustration. The second of these is a moderately high dimensional family, where the way that the boundaries of the simplex are attached to the model is of great importance for the behaviour of the likelihood and of the maximum likelihood estimate. In general, working in a simplex, boundary effects mean that standard first order asymptotic results can fail, while the much more flexible higher order methods can be very effective. The other example is a continuous curved exponential family, where both higher order asymptotic sampling theory results and geometrically-based dimension reduction are described.

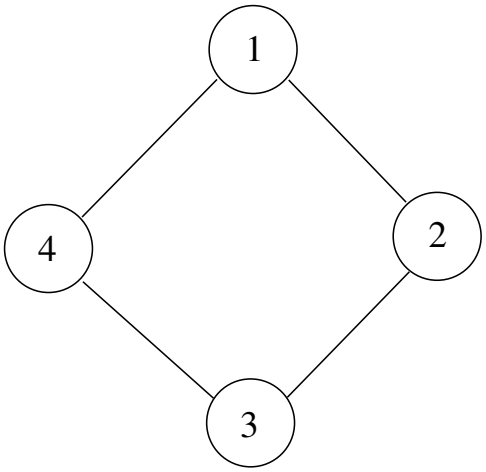
Example 2 *The paper [9] models survival times for leukaemia patients. These times, recorded in days, start at the time of diagnosis, and there are 43 observations; see [10] for details. We further assume that the data is censored at a fixed value. It was observed that a censored exponential distribution gives a reasonable, but not exact, fit. As discussed in [11], this gives a one-dimensional curved exponential family inside a two-dimensional regular exponential family of the form:*

$$\exp\left[\lambda_1 x + \lambda_2 y - \log\left\{\frac{1}{\lambda_2}\left(e^{\lambda_2 t} - 1\right) + e^{\lambda_1 + \lambda_2 t}\right\}\right], \tag{2}$$

where $y = \min(z,t)$ *and* $x = I(z \geq t)$, *and the embedding map is given by* $(\lambda_1(\theta), \lambda_2(\theta)) = (-\log\theta, -\theta)$.

As shown in [4], the loss due to discretisation can be made arbitrarily small for all information geometry objects. Thus, for example, using this computational approach, it is straightforward to compute the bias correction described in Example 1. Each of the terms in the asymptotic bias, i.e., the metric, g_{ij}, *its inverse,* g^{ij}, *the Christoffel symbols,* $\Gamma^{(-1)}_{ijk}$, *and curvature term,* $h^{(-1)}$, *can be directly numerically coded as appropriate finite difference approximations to derivatives. Thus, "black-box" code can directly calculate the numerical value of the asymptotic bias, and this numerical value can then be used by those who are not familiar with information geometry. For example this calculation establishes the fact that, with this particular data set, the sample size is such that the bias is inferentially unimportant.*

Figure 1. Undirected graphical model showing the cyclic graph of order four.



Example 3 *The paper [12] discusses an undirected graphical model based on the cyclic graph of order four, shown in Figure 1, with binary random variables at each node. Without any constraints, there are 16 possible values for the graph, so model space can be thought of as a* 15*-dimensional simplex, including the relative boundary. However, the conditional independence relations encoded by the graph impose linear constraints in the natural parameters of the exponential family. Thus, the resultant model is a lower dimensional full exponential family and its closure.*

As described in [12], the four cycle model is a seven dimensional exponential family, which is a $+1$*-affine subspace of the* $+1$*-affine structure of the* 15*-dimensional simplex. The model can be written in the form:*

$$\left(\frac{\pi_i \exp\left\{\sum_{h=1}^{8} \eta_h v_{hi}\right\}}{\sum_{j=0}^{15} \pi_j \exp\left\{\sum_{h=1}^{8} \eta_h v_{hj}\right\}} \right)_{i=0}^{15} \tag{3}$$

for a given set of linearly independent vectors $\{v_h\}_{h=1}^{8}$. *The existence of the maximum likelihood estimate for* $\eta = (\eta_h)$ *will depend on how the limit points of Model (3) meet the observed face of* Δ^{15}*; that is, the span of the vertices (bins) having positive counts. Thus, a key computational task is to learn how a full exponential family, defined by a representation of the form of (3), is attached to boundary sub-simplices of the high-dimensional embedding simplex.*

In order to visualise the geometric aspects of this problem, consider a lower dimensional version. Define a two-dimensional full exponential family by the vectors $v_1 = (1,2,3,4), v_2 = (1,4,9,-1)$ *and*

the uniform distribution base point, π_i, embedded in the three-dimensional simplex. The two-dimensional family is defined by the $+1$-affine space through $(0.25, 0.25, 0.25, 0.25)$ spanned by the space of vectors of the form:

$$\alpha(1, 2, 3, 4) + \beta(1, 4, 9, -1) = (\alpha + \beta, 2\alpha + 4\beta, 3\alpha + 9\beta, 4\alpha - \beta).$$

Consider directions from the origin obtained by writing $\alpha = \theta\beta$, giving, for each θ, a one-dimensional, full exponential family parameterized by β in the direction $\beta(\theta + 1, 2\theta + 4, 3\theta + 9, 4\theta - 1)$. The aspect of this vector, which determines the connection to the boundary, is the rank order of its elements. For example, suppose the first component was the maximum and the last the minimum. Then, as $\beta \to \pm\infty$, this one-dimensional family will be connected to the first and fourth vertex of the embedding four simplex, respectively. Note that changing the value of θ changes the rank structure, as illustrated in Figure 2. This plot shows the four element-wise linear functions of θ (dashed lines) and the salient overall feature of their rank order; that is, their upper and lower envelopes (solid lines). From this analysis of the envelopes of a set of linear functions, it can be seen that the function $2\theta + 4$ is redundant. The consequence of this is shown in Figure 3, which shows a direct computation of the two-dimensional family. It is clear that, indeed, only three of the four vertexes have been connected by the model.

In general, the problem of finding the limit points in full exponential families inside simplex models is a problem of finding redundant linear constraints. As shown in [13], this can be converted, via convex duality, into the problem of finding extremal points in a finite dimensional affine space. In the four-cycle model, this technique can construct all sub-simplices containing limit points of the four-cycle model. For example, it can be shown that all of the 16 vertices are part of the boundary. Once the boundary points have been identified as necessary and sufficient, conditions for the existence of the maximum likelihood in the $+1$-parameters can easily be found computationally [6].

Figure 2. The envelope of a set of linear functions. Functions, dashed lines; envelope, solid lines.

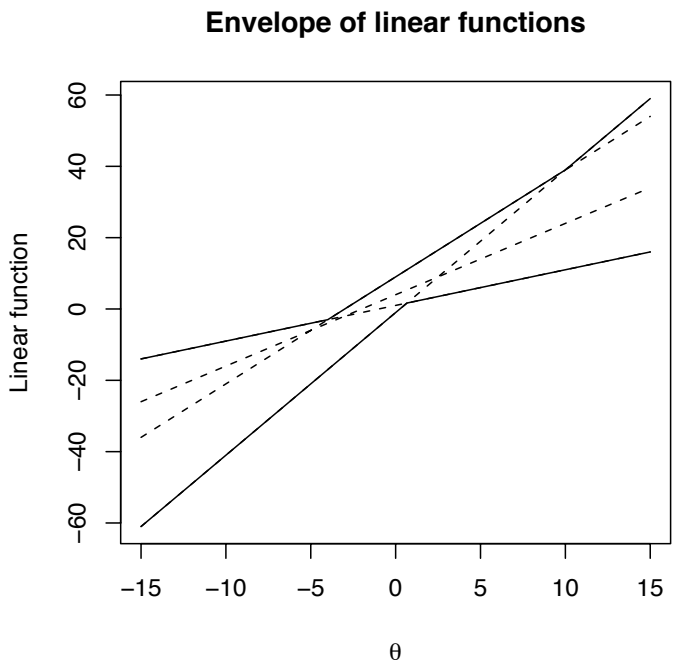
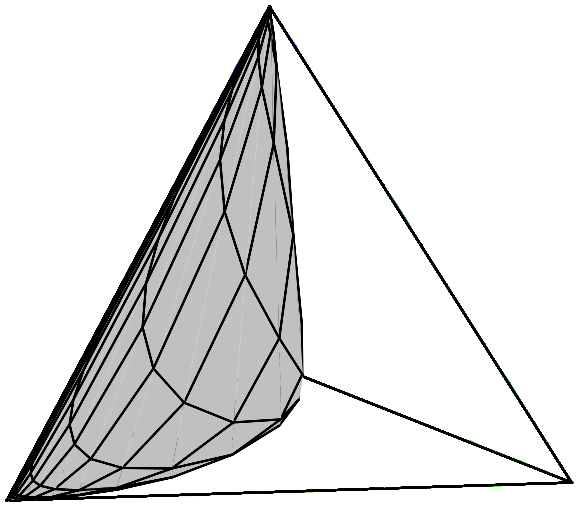


Figure 3. Attaching a two-dimensional example to the boundary of the simplex.



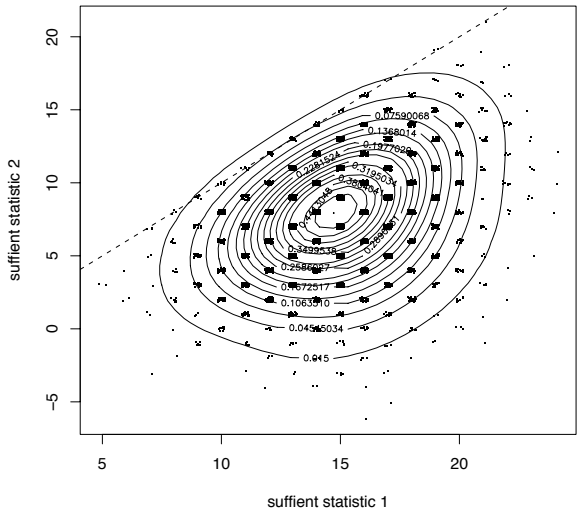
2.3. Tensor Analysis and Numerical Stability

One of the most powerful set of results from classical information geometry is the way that geometrically-based tensor analysis is perfect for use in multi-dimensional higher order asymptotic analysis; see [14] or [15]. The tensorial formulation does, however, present a couple of problems in practice. For many, its very tight and efficient notational aspects can obscure rather than enlighten, while the resulting formulae tend to have a very large number of terms, making them rather cumbersome to work with explicitly. These are not problems at all for the computational approach described in this paper. Rather, the clarity of the tensorial approach is ideal for coding, where large numbers of additive terms, of course, are easy to deal with.

Two more fundamental issues, which the global geometric approach of this paper highlights, concern numerical stability. The ability to invert the Fisher information matrix is vital in most tensorial formulae, and so understanding its spectrum, discussed in Section 2.4, is vital. Secondly, numerical underflow and overflow near boundaries require careful analysis, and so, understanding the way that models are attached to the boundaries of the extended multinomial models is equally important. The four-cycle model, to which we now return, illustrates computational information geometry doing this effectively.

Example 4 *The multivariate Edgeworth approximation to the sampling distribution of part of the sufficient statistic for the four-cycle model is shown in Figure 4. Using the techniques described above, a point near the boundary of the* 15*-simplex has been selected as the data generation process. For illustration, we focus on the marginal distribution of two components of the sufficient statistic, though any number could have been chosen. The boundary forces constraints on the range of the sufficient statistics, shown by the dashed line in the plot. The points, jittered for clarity, show the distribution computed by simulation. It is typical that such boundary constraints prevent standard first order methods from performing well, but the greater flexibility of higher order methods can be seen to work well here. As discussed above, methods, such as the multivariate Edgeworth expansion, can be strongly exploited in a computational framework, such as ours. Note, the discretization that can be observed in the figure is extensively discussed in [6].*

Figure 4. Using the Edgeworth expansion near the boundary of four-cycle model.



2.4. Spectrum of Fisher Information

We focus now on the second numerical issue identified above. In any multinomial, the Fisher information matrix and its inverse are explicit. Indeed, the 0-geodesics and the corresponding geodesic distance are also explicit; see [2] or [3]. However, since the simplex glues together multinomial structures with different supports and the computational theory is in high dimensions, it is a fact that the Fisher information matrix can be arbitrarily close to being singular. It is therefore of central interest that the spectral decomposition of the Fisher information itself has a very nice structure, as shown below.

Example 5 *Consider a multinomial distribution based on* 81 *equal width categories on* $[-5, 5]$*, where the probability associated to a bin is proportional to that of the standard normal distribution for that bin. The Fisher information for this model is an* 80×80 *matrix, whose spectrum is shown in Figure 5. By inspection, it can be seen that there are exponentially small eigenvalues, so that while the matrix is positive definite, it is also arbitrarily close to being singular. Furthermore, it can be seen that the spectrum has the shape of a half-normal density function and that the eigenvalues seem to come in pairs. These facts are direct consequences of the general results below.*

With π_{-0} denoting the vector of all bin probabilities, except π_0, we can write the Fisher information matrix (in the $+1$ form) as N times:

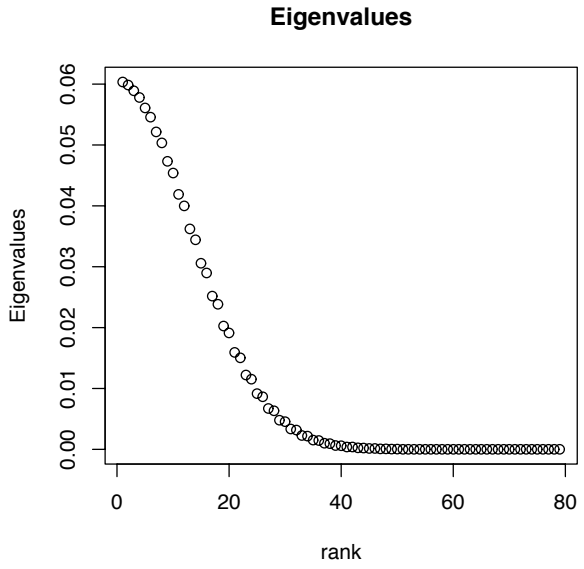
$$I(\pi) := diag(\pi_{-0}) - \pi_{-0}\pi_{-0}^T.$$

This has an explicit spectral decomposition, which can be computed by using interlacing eigenvalue results (see for example [16], Chapter 4). In particular, if the diagonal matrices, $diag(\pi_1, \ldots, \pi_k)$ and $diag(\lambda_1 I_{m_1} | \cdots | \lambda_g I_{m_g})$, agree up to a row-and-column permutation, where $g > 1$ and $\lambda_1 > \cdots > \lambda_g > 0$, then $I(\pi)$ has ordered spectrum:

$$\lambda_1 > \tilde{\lambda}_1 > \cdots > \lambda_g > \tilde{\lambda}_g \geq 0, \tag{4}$$

with $\tilde{\lambda}_g > 0 \iff \pi_0 > 0$, each λ_i having multiplicity $m_i - 1$, while each $\tilde{\lambda}_g$ is simple.

Figure 5. Spectrum of the Fisher information matrix of a discretised normal distribution.



We give a complete account of the spectral decomposition (SpD) of $I(\pi)$. There are four cases to consider, the last having the generic spectrum of (4). Without loss, after permutation, assume now $\pi_1 \geq \cdots \geq \pi_k$. The four cases are:

Case 1 For some $l < k$, the last $k - l$ elements of π_{-0} vanish: the sub-case $l = 0 \iff \pi_0 = 1 \iff I(\pi) = 0$ is trivial. Otherwise, writing $\pi_+ = (\pi_1, \ldots, \pi_l)^T$ and $\Pi_+ = diag(\pi_+)$, the SpD of:

$$I(\pi) = \left(\begin{array}{c|c} \Pi_+ - \pi_+ \pi_+^T & 0 \\ \hline 0 & 0 \end{array} \right)$$

follows at once from that of $\Pi_+ - \pi_+ \pi_+^T$, given below.

Case 2 $k = 1$: this case is trivial.

Case 3 $k > 1, \pi = \lambda 1_k, \lambda > 0$: the SpD of $I(\pi)$ is:

$$\lambda C_k + \lambda(1 - k\lambda) J_k$$

where $C_k = I_k - J_k$ and $J_k = k^{-1} 1_k 1_k^T$. Here, λ has multiplicity $k - 1$ and eigenspace $[Span(1_k)]^{\perp}$, while $\tilde{\lambda} := \lambda(1 - k\lambda)$ has multiplicity one and eigenspace $Span(1_k)$. In particular, since $1 - \pi_0 = k\lambda$, it follows that:

$$I(\pi) \text{ is singular } \iff \pi_0 = 0.$$

Case 4 $\pi_{-0} = (\lambda_1 1_{m_1}^T | \ldots | \lambda_g 1_{m_g}^T)^T$, $g > 1$ and $\lambda_1 > \cdots > \lambda_g > 0$:

This is the generic case, having the spectrum of (4) above. Denoting by O_m the zero matrix of order $m \times m$ and by $P(\nu)$ the rank one orthogonal projector onto Span(ν), ($\nu \neq 0$), the SpD is:

$$\sum_{i=1, m_i > 1}^{g} \lambda_i diag(O_{m_{i-}}, C_{m_i}, O_{m_{i-}}) + \sum_{i=1}^{g} \tilde{\lambda}_i P \left(\left(\frac{\lambda_1}{\tilde{\lambda}_i - \lambda_1} 1_{m_1}^T, \ldots, \frac{\lambda_g}{\tilde{\lambda}_i - \lambda_g} 1_{m_g}^T \right)^T \right),$$

where: $m_{i-} = \sum\{m_j | j < i\}$, $m_{i+} = \sum\{m_j | j > i\}$ and the $\tilde{\lambda}_i$ are the zeros of:

$$h(\tilde{\lambda}) := 1 + \sum_{i=1}^{g} \frac{m_i \lambda_i^2}{\tilde{\lambda} - \lambda_i} = (1 - \sum_{i=1}^{g} m_i \lambda_i) + \tilde{\lambda} \left(\sum_{i=1}^{g} \frac{m_i \lambda_i}{\tilde{\lambda} - \lambda_i} \right).$$

In particular, $\{\tilde{\lambda}_i : i = 1, \cdots, g\}$ are simple eigenvalues satisfying (4) while, whenever $m_i > 1$, λ_i, is also an eigenvalue having multiplicity $m_i - 1$. Further, expanding $\det(I(\pi))$, we again find:

$$I(\pi) \text{ is singular } \iff \pi_0 = 0,$$

so that $\widetilde{\lambda}_g > 0 \Leftrightarrow \pi_0 > 0$, as claimed. Finally, we note that each $\widetilde{\lambda}_i$ $(i < g)$ is typically (much) closer to λ_i than to λ_{i+1}. For, considering the graph of $x \to 1/x$, $h\left((\lambda_i + \lambda_{i+1})/2 + \delta(\lambda_i - \lambda_{i+1})/2\right)$ $(-1 < \delta < +1)$ is well-approximated by:

$$1 - \frac{2m_i\lambda_i^2}{(\lambda_i - \lambda_{i+1})(1 - \delta)} + \frac{2m_{i+1}\lambda_{i+1}^2}{(\lambda_i - \lambda_{i+1})(1 + \delta)}$$

whose unique zero δ_* over $(-1, 1)$ is positive whenever, as will typically be the case, $m_i = m_{i+1}$ (both will usually be one), while $(m_i\lambda_i + m_{i+1}\lambda_{i+1}) < 1/2$. Indeed, a straightforward analysis shows that, for any m_i and m_{i+1}, $\delta_* = 1 + O(\lambda_i)$ as $\lambda_i \to 0$.

2.5. Total Positivity and Local Mixing

Mixture modelling is an exemplar of a major area of statistics in which computational information geometry enables distinctive methodological progress. The -1-convex hull of an exponential family is of great interest, mixture models being widely used in many areas of statistical science. In particular, they are explored further in [5]. Here, we simply state the main result, a simple consequence of the total positivity of exponential families [17], that, generically, convex hulls are of maximal dimension. In this result, "generic" means that the $+1$ tangent vector, which defines the exponential family as having components that are all distinct.

Theorem 1 *The -1-convex hull of an open subset of a generic one-dimensional exponential family is of full dimension.*

Proof 1 *For any $(\pi_i) \in \Delta^k$ with each $\pi_i > 0$, $\theta_0 < \cdots < \theta_k$ and $s_0 < \cdots < s_k$, let $B = (\pi(\theta_0), ..., \pi(\theta_k))$ have general element:*

$$\pi_i(\theta_j) := \pi_i \exp[s_i\theta_j - \psi(\theta_j)].$$

Further, let $\widetilde{B} = B - \pi(\theta_0)1_{k+1}^T$, whose general column is $\pi(\theta_j) - \pi(\theta_0)$. Then, it suffices to show that $\widetilde{B}$ has rank k. However, using [18] (p. 33), $Rank(\widetilde{B}) = Rank(B) - 1$, so that:

$$Rank(\widetilde{B}) = k \Leftrightarrow B \text{ is nonsingular } \Leftrightarrow B^* \text{ is nonsingular,}$$

where $B^* = (\exp[s_i\theta_j])$. It suffices, then, to recall [17] that $K(x, y) = \exp(xy)$ is strictly total positive (of order ∞), so that $\det B^* > 0$.

3. Infinite Dimensional Structure

This section will start to explore the question of whether the simplex structure, which describes the finite dimensional space of distributions, can extend to the infinite dimensional case. We examine some of the differences with the finite dimensional case, illustrating them with clear, commonly occurring examples.

3.1. Infinite Dimensional Information Geometry: A Review

In the previous sections, the underlying computational space is always finite dimensional. This section looks at issues related to an infinite dimensional extension of the theory in that paper. There is a great deal of literature concerning infinite dimensional statistical models. The discussion here concentrates on information geometric, parametrisation and boundary issues.

The information geometry theory of Amari [3] has a geometric foundation, where statistical models (typically full and curved exponential families) have a finite dimensional manifold structure. When considering the extension to infinite dimensional cases, Amari notes the problem of finding an "adequate topology" [3] (p. 93). There has to be very interesting work following up this topological challenge. By concentrating on distributions with a common support, the paper [19] uses the geometry of a Banach manifold, where local patches on the manifold are modelled by Banach spaces, via the concept of an Orlicz space. This gives a structure that is analogous to an infinite dimensional exponential family, with mean and natural parameters and including the ability to define mixed parametrisations. One drawback of this Banach structure, as pointed out in [20], is that the likelihood function with finite samples is not continuous on the manifold. Fukumizu uses a reproducing kernel Hilbert space structure rather than a Banach manifold, which is a stronger topology. There are strong connections between the approach taken in [20] and the material in Section 3.2, we note two issues here: (1) a focus on the finite nature of the data; and (2) using a Hilbert structure defined by a cumulant generating function. The approaches differ in that [20] uses a manifold approach rather than the simplicial complex as the fundamental geometric object. There is also other work that explicitly used infinite dimensional Hilbert spaces in statistics, a good reference being [21].

In this paper, in contrast to previous authors, a simplicial, rather than a manifold-based, approach is taken. This allows distributions with varying support, as well as closures of statistical families to be included in the geometry. Another difference in approach is the way in which geometric structures are induced by infinite dimensional affine spaces rather than by using an intrinsic geometry. This approach was introduced by [7] and extended by [8]. Spaces of distributions are convex subsets of the affine spaces, and their closure within the affine space is key to the geometry.

In exponential families, the -1-affine structure is often called the mean parametrisation, and using moments as parameters is one very important part of modelling. In the infinite dimensional case, the use of moments as a parameter system is related to the classical moment problem—when does there exist a (unique) distribution whose moments agree with a given sequence?—which has generated a vast literature in its own right; see [22–24]. In general terms, the existence of a solution to the moment problem is connected to positivity conditions on moment matrices. Such conditions have been used in connection to the infinite dimensional geometry of mixture models [25]. Uniqueness, however, is a

much more subtle problem: sufficient conditions can be formulated in terms of the rate of growth of the moments [24]. Counter examples to general uniqueness results include the log-normal distribution [23].

The geometry of the Fisher information is also much more complex in general spaces of distributions than in exponential families. Simple mixture models, including two-component mixtures of exponential distributions [26], can have "infinite" expected Fisher information, which gives rise to non-standard inference issues. Similar results on infinitely small (and large) eigenvalues of covariance operators are also noted in [20]. Since the Fisher information is a covariance, the fact that it does not exist for certain distributions or that its spectrum can be unbounded above or arbitrarily close to zero is not a surprise. However, these observations do need to be taken into account when considering the information geometry of infinite dimensional spaces.

The rest of this section looks at the topology and geometry of the infinite dimensional simplex and gives some illustrative examples, which, in particular, show the need for specific Hilbert space structures, discussed in the final section.

3.2. Topology

For simplicity and concreteness, in this section, we will be looking at models for real valued random variables. In this paper, we restrict attention to the cases where the sample space is $\mathbb{R}^+$ or $\mathbb{R}$ and has been discretised to a countably infinite set of bins, B_i, with $i \in \mathbb{N}$ or $\mathbb{Z}$, respectively. In the finite case, the basic object is the standard simplex, Δ^k, with $k+1$ bins. We generalise this to countable unions of such objects. Of these, one is of central importance, denoted by Δ_{emp} or simply Δ, because it is the smallest object that contains all possible empirical distributions.

Definition 2 *For any finite subset of bins, indexed by $\mathcal{I} \subset \mathbb{N}$ or $\mathbb{Z}$, denote*

$$\Delta_{\mathcal{I}} = \left\{ \mathbf{x} = (x_i)_{i \in \mathcal{I}} : x_i \geq 0 \, , \sum_{i \in \mathcal{I}} x_i = 1 \right\} .$$

We take the union of all such sets $\bigcup_{|\mathcal{I}|<\infty} \Delta_{\mathcal{I}}$, where $|\mathcal{I}|$ denotes the number of elements of the index set. This can always be written as:

$$\Delta = \left\{ \mathbf{x} = (x_i)_{i \in \mathbb{Z}} : \sum_{i \in \mathbb{Z}} x_i = 1, x_i \geq 0 \text{ and only finitely many } x_i > 0 \right\} .$$

In what follows, it is important to note that for any given statistical inference problem, the sample size, n, is always finite, even if we frequently use asymptotic approximations, where $n \to \infty$. Thus, the data, as represented by the empirical distribution, naturally lie in the space, Δ. However, many models, used in the given inference problem, will have support over all bins, so the models most naturally lie in the "boundary" constructed using the closures of the set. These objects are subsets of sequence spaces, and the corresponding topologies can be constructed from the Banach spaces, ℓ_p, $p \in [1, \infty]$. The following results follow directly from explicit calculations, where we note that in this section, since all terms are non-negative, convergence always means absolute convergence. In particular, arbitrary rearrangements of series do not affect the existence of limits or their values.

Example 6 *Consider the sequence of "uniform distributions"* $\mathbf{x}^{(n)} = (\frac{1}{n}, \dots, \frac{1}{n}, 0, \dots)$ *as elements of* Δ. *This has an* ℓ_p *limit of the zero sequence for* $p \in (1, \infty]$.

Proposition 1 *The* ℓ_p *extreme points of* Δ, *for* $p \in (1, \infty]$, *are the zero sequence and the sequences,* δ_i $(i \in \mathbb{Z})$, *with one as the* $i-th$ *element and zero elsewhere.*

For $p \in [1, \infty]$, let $\overline{\Delta}_p \subset \ell_p$ denote the ℓ_p closure of Δ.

Theorem 2 *(a)* $\overline{\Delta}_1 = \left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathcal{I}} x_i = 1\right\}$.
(b) $\overline{\Delta}_\infty = \left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathcal{I}} x_i \leq 1\right\}$.
(c) For $p \in (1, \infty)$, $\overline{\Delta}_p = \overline{\Delta}_\infty = \left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathcal{I}} x_i \leq 1\right\}$.

Proof 2 *(a) It is immediate that* $\left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathbb{Z}} x_i = 1\right\} \subseteq \overline{\Delta}_1$. *Conversely, if* $\bar{x}$ *is a limit point, then all its elements must be non-negative. Finally, if* $\sum_{i=1}^{\infty} \bar{x}_i$ *is not bounded above by one, then there exists* N, *such that* $\sum_{i=1}^{N} \bar{x}_i > 1 + \epsilon$ *for some* $\epsilon > 0$. *Hence,* $\sum_{i=1}^{\infty} |\bar{x}_i - x_i^{(n)}| \geq \sum_{i=1}^{N} |\bar{x}_i - x_i^{(n)}| \geq \sum_{i=1}^{N} \bar{x}_i - \sum_{i=1}^{N} x_i^{(n)} > \epsilon$ *for all* n, *which contradicts convergence. If* $\sum_{i=1}^{\infty} \bar{x}_i < 1 - \epsilon$, *then* $\sum_{i=1}^{\infty} |\bar{x}_i - x_i^{(n)}| \geq \sum_{i=1}^{\infty} x_i^{(n)} - \sum_{i=1}^{\infty} \bar{x}_i > \epsilon$, *which again contradicts convergence.*

(b) It is again immediate that $\left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathbb{Z}} x_i = 1\right\} \subseteq \overline{\Delta}_\infty$. *However, by Example 6, the zero sequence is also in* $\overline{\Delta}_\infty$, *so that* $\left\{\mathbf{x} = (x_i)_{i\in\mathbb{Z}} : x_i \geq 0, \sum_{i\in\mathbb{Z}} x_i \leq 1\right\} \subseteq \overline{\Delta}_\infty$.

Conversely, by contradiction, it is easy to see that all elements of the closure must have non-negative elements. Finally, for any $\bar{x} \in \overline{\Delta}_\infty$, *if* $\sum_{i=1}^{\infty} \bar{x}_i$ *is not bounded above by one, there exists* N, *such that* $\sum_{i=1}^{N} \bar{x}_i > 1 + \epsilon$ *for some* $\epsilon > 0$. *For any sequence of points,* $x^{(n)}$ *in* Δ, *we have that* $\sum_{i=1}^{N} x_i^{(n)} \leq 1$, *so that, for* $i = 1, \dots, N$, *the maximum value of* $|x_i^{(n)} - \bar{x}_i| > \epsilon/N$. *Hence, for all sequences,* $x^{(n)}$, *we have* $\|x^{(n)} - \bar{x}\|_\infty > \epsilon/N$, *which contradicts* $\bar{x}$ *being in the closure.*

(c) This follows essentially the same argument as (b) by noting in the case where $\sum_{i=1}^{\infty} \bar{x}_i$ *is not bounded above by one, we have:*

$$\|x^{(n)} - \bar{x}\|_p^p \geq \sum_{i=1}^{N} |\bar{x}_i - x_i^{(n)}|^p \geq N \max_{i=1,\dots N} |x_i^{(n)} - \bar{x}_i|^p > N^{1-p}\epsilon^p$$

for any sequences, $x^{(n)}$, *which contradicts* $\bar{x}$ *being in the closure.*

It is immediate that the spaces, Δ and $\overline{\Delta}_1$, are convex subsets of ℓ_1 and that $\overline{\Delta}_\infty$ is a convex set in ℓ_∞.

3.3. Geometry

In the same way as for the finite case, the -1-geometry can be defined using an affine space structure using the following definition.

Definition 3 *Let* $\mathcal{I}$ *be a countable index set which is a subset of* $\mathbb{Z}$. *The* -1*-affine space structure over distributions is* $(X_{mix}, V_{mix}, +)$, *where:*

$$X_{mix} = \left\{\mathbf{x} = (x_i) | \sum_{\mathcal{I}} x_i = 1, \sum_{\mathcal{I}} |x_i| < \infty\right\}, V_{mix} = \left\{\mathbf{v} = v_i | \sum_{\mathcal{I}} v_i = 0, \sum_{\mathcal{I}} |v_i| < \infty, \right\},$$

and $\mathbf{x} + \mathbf{v} = (x_i + v_i)$.

In order to define the $+1$-geometric structure, we also follow the approach used in the finite case. Initially, to understand the $+1$- structure, consider the case where all distributions have a common support, *i.e.*, assume $\pi_i > 0$ for all i. We follow here the approach of [7].

Definition 4 *Consider the set of non-negative measures on $\mathbb{N}$ or $\mathbb{Z}$ and the equivalence relation defined by:*

$$\{a_i\} \sim \{b_i\} \iff \exists \lambda > 0 \, s.t. \, \forall i \; a_i = \lambda b_i.$$

The equivalences classes of this are the points in the $+1$ geometry.

These points can be further partitioned into sets with the same support, i.e., $\mathrm{supp}(< a >) = \{i : a_i > 0\}$, where this is clearly well-defined.

On sets of $+1$-points with the same support, we can define the $+1$-geometry in the same way as in the finite case. With "exp" connoting an exponential family distribution, we have:

Definition 5 *For a given index set, $\mathcal{I}$, define X_{exp} to be all $+1$-points whose support equals $\mathcal{I}$, and define the vector space $V_{exp} = \{v_i, i \in \mathcal{I}\}$ with the operation, $\oplus$, defined by:*

$$< x_i > \oplus v_i = \langle x_i \exp(v_i) \rangle ,$$

is an affine space. The $+1$-affine structure is then defined by $(X_{exp}, V_{exp}, \oplus)$.

Theorem 3 *If* $\mathbf{a}$ *and* $\mathbf{b}$ *lie in Δ (or $\overline{\Delta}_1$) and have the same support, then $C(\rho) = \sum (a_i^\rho b_i^{(1-\rho)}) < \infty$ for $\rho \in [0, 1]$. Hence, $\frac{a_i^\rho b_i^{(1-\rho)}}{C(\rho)} \in \Delta$ (or $\overline{\Delta}_1$).*

Proof 3 *Since a, b are absolutely convergent, the sequence, $\max(a_i, b_i)$, is also. Since we have:*

$$0 \leq \min(a_i, b_i) \leq a_i^\rho b_i^{1-\rho} \leq \max(a_i, b_i)$$

it follows that $C(\rho) < \infty$, and we have the result.

This result shows that sets in $\overline{\Delta}_1$ with the same support are $+1$-convex, just as the faces in the finite case are.

3.4. Examples

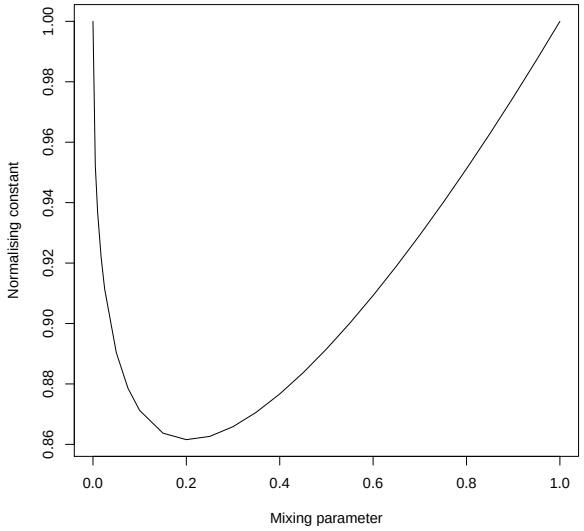
In order to get a sense of how the $+1$-geometry works, let us consider a few illustrative examples.

Example 7 *If we denote the discretised standard normal density by* $\mathbf{a}$ *and the discretised Cauchy density by* $\mathbf{b}$ *and consider the path:*

$$\frac{a_i^\rho b_i^{(1-\rho)}}{C(\rho)},$$

the normalising constant is shown in Figure 6. We see that at $\rho = 0$ (the Cauchy distribution), we have that the derivative of the normalising constant (i.e., the mean of the sufficient statistic) is tending to infinity. At the other end ($\rho = 1$), the model can be extended in the sense that the distribution exists for values greater than one.

Figure 6. Normalising constant for normal-Cauchy exponential mixing example.



Thus, in this example, the path joining the two distributions is an extended, rather than natural, exponential family, since we have to include the boundary point where the mean is unbounded.

Example 8 *Let us return to Example 2, but now without the censoring. Thus, now, there is a countably infinite set of bins, and so, we can investigate its embedding in the infinite simplex. As discussed in [4], we shall discretise the continuous distribution by computing the probabilities associated to bins* $[c_i, c_{i+1}]$, $i = 1, 2, \cdots$.

For the exponential model, $Exp(\theta)$*, the bin probabilities are simply:*

$$\pi_i(\theta) = \exp(-\theta c_i) - \exp(-\theta c_{i+1}).$$

Using this, the model will lie in the infinite simplex on the positive half line with the index set $\mathcal{I} = \mathbb{N}$.

First, consider the case where we have a uniform choice of discretisation, where $c_n = n \times \epsilon$ *for some fixed,* $\epsilon > 0$*. In this case, the bin probabilities can be written as an exponential family:*

$$\pi_n(\theta) = \exp\left[-\theta\epsilon n + \log(1 - e^{-\theta\epsilon})\right]$$

for $\theta > 0$*. This gives a* $+1$*-geodesic though* $\{\pi_i(\theta_0)\}$ *in the direction* $\{\epsilon \times n\}$ *of the form:*

$$\pi_n(\theta_0) \exp\left[-\lambda\epsilon n + \log\left(\frac{1 - e^{-(\lambda+\theta_0)\epsilon}}{1 - e^{-\theta_0\epsilon}}\right)\right] \tag{5}$$

for $\lambda > -\theta_0$*. In the case where* $\lambda \to -\theta_0$*, the limiting distribution is the zero measure in* $\overline{\Delta}_\infty$*, and at the other extreme, where* $\lambda \to \infty$*, the limiting distribution is the atomic distribution in the first bin, a distribution with a different support than* $\pi_i(\theta_0)$*. However, unlike the finite case, there is no guarantee that, for a given "direction",* $\{t_i\}$*, there exists a* $+1$*-geodesic starting at* $\{\pi_i(\theta_0)\}$*, since we require the convergence of the normalising constant:*

$$\sum_{i=0}^{\infty} \pi_i(\theta_0) \exp(\lambda t_i) < \infty.$$

From this example, we see that the limit points of exponential families can lie in the space, $\overline{\Delta}_\infty$, but not in $\overline{\Delta}_1$. The next example shows that limits do not have to exist at all.

Example 9 *Consider the family whose bin probabilities, $\pi_i \in \overline{\Delta}_\infty$, are proportional to a discretised standard normal with bins of constant width. The exponential family, which is proportional to $\pi_i \exp(\theta i)$, does not have an ℓ_∞ limit, as it is discretised normal with mean θ. The natural parameter space here is $(-\infty, \infty)$.*

The last illustrative example is from [26] and shows that even for simple models, the Fisher information for the parameters of interest need not be finite.

Example 10 *Let us consider a simple example of a two-component mixture of (discretised) exponential distributions:*

$$(1-\rho)\pi_i(\theta_0+\lambda) + \rho\pi_i(\theta_0) \tag{6}$$

the tangent vector in the ρ-direction is:

$$\pi_i(\theta_0) - \pi_i(\theta_0+\lambda) = \pi_i(\theta_0)\left(1 - e^{-\lambda\epsilon n}C\right)$$

for a positive constant, C. The corresponding squared length, with respect to the Fisher information, is:

$$\sum_{n=0}^{\infty} \frac{\left(1-e^{-\lambda\epsilon n}C\right)^2}{\pi_i(\theta_0)}.$$

As an example, consider $\theta_0 = 1$; then, this term will be infinite for $\lambda \leq -0.5$.

3.5. Hilbert Space Structures

Following these examples, we can consider the Hilbert space structure of exponential families inside the infinite simplex with the following results.

Definition 6 *Define the functions, $S(\cdot)$, by $S(\{v_i\}, \pi) = \sup_\theta \{\theta | \sum_{\mathcal{I}} \pi_i \exp(\theta v_i) < \infty\}$, the function being set to ∞ when the set is unbounded. Furthermore, define for a given $\{\pi_i\} \in \bar{\Delta}_\infty$, the set:*

$$V(\pi) = \{\{v_i\} | S(\{v_i\}, \pi > 0\}\,, \text{ and } V^c(\pi) = \{\{v_i\} | \pm\{v_i\} \in V(\pi)\}\,.$$

The spaces, $V^c(\pi)$, correspond to the directions in which the $+1$-geodesic and, so, the corresponding exponential families are well-defined and have particularly "nice" geometric structures.

Theorem 4 *For π, define a Hilbert space by:*

$$H(\pi) := \left\{\{v_i\} | \sum v_i^2 \pi_i < \infty\right\}$$

with inner product:

$$\langle\{v_i\}, \{w_i\}\rangle_\pi = \sum v_i w_i \pi_i,$$

and corresponding norm $||\cdot||_\pi$. Under these conditions:
(i) $V^c(\pi)$ is a subspace of $H(\pi)$, and
(ii) the set $V(\pi)$ is a convex cone.

Proof 4 *(i) First, if* $\{v_i\} \in V^c(\pi)$*, then by definition, the moment generating function:*

$$\sum \exp(\theta v_i)\pi_i,$$

is finite for θ *in an open set containing* $\theta = 0$*. Hence, have both:*

$$\sum v_i\pi_i < \infty, \text{ and } \sum v_i^2\pi_i < \infty.$$

Thus, $\{v_i\} \in H(\pi)$*. The fact that it is a subspace follows from (ii) below.*

(ii) It is immediate that $V(\pi)$ *is a cone.*

Convexity follows from the Cauchy–Schwartz inequality, since for all $\{v_i\}, \{v_i^*\} \in V(\pi)$ *and* $\lambda \in [0, 1]$*, it follows that:*

$$\begin{aligned} \left\{\sum \pi_i e^{\frac{\theta}{2}(\lambda v_i + (1-\lambda)v_i^*)}\right\}^2 &= \left\{\sum \left(\sqrt{\pi_i}e^{\frac{\theta}{2}\lambda v_i}\right)\left(\sqrt{\pi_i}e^{\frac{\theta}{2}(1-\lambda)v_i^*}\right)\right\}^2 \\ &\leq \left\{\sum \pi_i e^{\theta\lambda v_i}\right\}\left\{\sum \pi_i e^{\theta(1-\lambda)v_i^*}\right\}, \end{aligned}$$

and, so, is finite for a strictly positive value of θ*, hence* $\{\lambda v_i + (1-\lambda)v_i^*\} \in V(\pi)$.

Hence, this result illustrates the point above regarding the existence of "nice" geometric structure in the sense of Amari's information geometry developed for finite dimensional exponential families. Infinite dimensional families have a richer structure; for example, they include the possibility of having an infinite Fisher information; see Examples 7 and 10.

Acknowledgments

The authors would like to thank Karim Anaya-Izquierdo and Paul Vos for many helpful discussions and the UK's Engineering and Physical Sciences Research Council (EPSRC) for the support of grant number EP/E017878/.

Author Contributions

All authors contributed to the conception and design of the study, the collection and analysis of the data and the discussion of the results. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Efron, B. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.* **1975**, *3*, 1189–1242.
2. Kass, R.E.; Vos, P.W. *Geometrical Foundations of Asymptotic Inference*; John Wiley & Sons: London, UK, 1997.
3. Amari, S.-I. *Differential-Geometrical Methods in Statistics*; Lecture Notes in Statistics; Springer-Verlag Inc.: New York, NY, USA, 1985; Volume 28.

4. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P.; Vos, P. Computational Information Geometry: Foundations. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8085, pp. 311–318.
5. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P.; Vos, P. Computational Information Geometry: Mixture Modelling. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8085, pp. 319–326.
6. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P. When are first order asymptotics adequate? A diagnostic. *Stat* **2014**, *3*, 17–22.
7. Murray, M.K.; Rice, J.W. *Differential Geometry and Statistics*; Chapman & Hall: London, UK, 1993.
8. Marriott, P. On the local geometry of mixture models. *Biometrika* **2002**, *89*, 95–97.
9. Hand, D.J.; Daly, F.; Lunn, A.D.; McConway, K.J.; Ostrowski, E. *A Handbook of Small Data Sets*; Chapman and Hall: London, UK, 1994.
10. Bryson, M.C.; Siddiqui, M.M. Survival times: Some criteria for aging. *J. Am. Stat. Assoc.* **1969**, *64*, 1472–1483.
11. Marriott, P.; West, S. On the geometry of censored models. *Calcutta Stat. Assoc. Bull.* **2002**, *52*, 567–576.
12. Geiger, D.; Heckerman, D.; King, H.; Meek, C. Stratified exponential families: Graphical models and model selection. *Ann. Stat.* **2001**, *29*, 505–529.
13. Edelsbrunner, H. *Algorithms in Combinatorial Geometry*; Springer-Verlag: NewYork, NY, USA, 1987.
14. Barndorff-Nielsen, O.E.; Cox, D.R. *Asymptotic Techniques for Use in Statistics*; Chapman & Hall: London, UK, 1989.
15. McCullagh, P. *Tensor Methods in Statistics*; Chapman & Hall: London, UK, 1987.
16. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge Universtiy Press: Cambridge, UK, 1985.
17. Karlin, S. *Total Positivity*; Stanford University Press: Stanford, CA, USA, 1968; Volume I.
18. Householder, A.S. *The Theory of Matrices in Numerical Analysis*; Dover Publications: Dover, DE, USA, 1975.
19. Pistone, G.; Rogantin, M.P. The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli* **1999**, *5*, 571–760.
20. Fukumizu, K. Infinite dimensional exponential families by reproducing kernel Hilbert spaces. In Proceedings of the 2nd International Symposium on Information Geometry and its Applications, Tokyo, Japan, 12–16 December 2005.
21. Small, C.G.; McLeish, D.L. *Hilbert Space Methods in Probability and Statistical Inference*; John Wiley & Sons: London, UK, 1994.
22. Akhiezer, N.I. *The Classical Moment Problem*; Hafner: New York, NY, USA, 1965.
23. Stoyanov, J.M. *Counter Examples in Probability*; John Wiley & Sons: London, UK, 1987.
24. Gut, A. On the moment problem. *Bernoulli* **2002**, *8*, 407–421.
25. Lindsay, B.G. Moment matrices: Applications in mixtures. *Ann. Stat.* **1989**, *17*, 722–740.

26. Li, P.; Chen, J.; Marriott, P. Non-finite Fisher information and homogeneity: An EM approach. *Biometrika* **2009**, *96*, 411–426.